CommunityKG-RAG: Leveraging Community Structures in Knowledge Graphs for Advanced Retrieval-Augmented Generation in Fact-Checking

Anonymous EMNLP submission

Abstract

Despite advancements in Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) systems, their effectiveness is often hindered by a lack of integration with entity relationships and community structures, limiting their ability to provide contextually rich and accurate information retrieval for fact-checking. We introduce CommunityKG-RAG (Community Knowledge Graph-Retrieval Augmented Generation), a novel zero-shot framework that integrates community structures within Knowledge Graphs (KGs) with RAG systems to enhance the fact-checking process. Capable of adapting to new domains and queries without additional training, CommunityKG-RAG utilizes the multi-hop nature of community structures within KGs to significantly improve the accuracy and relevance of information retrieval. Our experimental results demonstrate that CommunityKG-RAG outperforms traditional methods, representing a significant advancement in fact-checking by offering a robust, scalable, and efficient solution.

1 Introduction

The occurrence of misinformation and the imperative of fact-checking are pivotal elements within the digital information ecosystem, profoundly affecting public discourse and shaping societal decisions worldwide. Concurrently, the advent of Large Language Models (LLMs) has unveiled remarkable capabilities in comprehending and producing human languages, presenting a promising avenue for bolstering fact-checking endeavors. Prior research (Buchholz, 2023; Li et al., 2023b; Caramancion, 2023; Hoes et al., 2023; Huang and Sun, 2023) has delved into directly prompting LLM models to identify false information. However, while LLMs can be instrumental in combating misinformation, their practical application still exposes two critical limitations. Firstly, these models are constrained by the cut-off date of their training data. Secondly, this issue is compounded by the tendency of LLMs to generate incorrect information or "hallucinations" (Huang et al., 2023) which could jeopardize the accuracy of claim verification in fact-checking tasks.

In response to these challenges, Retrieval-Augmented Generation (RAG) has emerged as a promising approach. By integrating the generative capabilities of LLMs with external data retrieval, RAG significantly enhances the accuracy and relevance of the responses. For instance, Liao et al. (2023) leverages RAG by employing both the dot product and the BERT-based sequence tagging model to identify key evidences. Soleimani et al. (2019) uses the BERT model to retrieve and validate claims.

While RAG significantly advances the capabilities of LLMs, it, too, faces unique challenges. Firstly, language models suffer from utilizing contexts in long texts. When crucial information is located in the middle, it is less likely to be effectively utilized by language models (Liu et al., 2023). Secondly, when contexts are laden with noise or contradictory information, RAG's performance can be adversely underscored (Barnett et al., 2024). Thirdly, the retrieval process plays a crucial role. Often, even if the answer to a query is present in the document corpus, it may not rank highly enough to be returned to the user (Barnett et al., 2024). Further expanding on the challenges in RAG systems, knowledge retrieved by these systems does not always contribute positively (Wang et al., 2023) and can sometimes detrimentally impact the original responses generated by the LLMs.

Acknowledging the challenges inherent in RAG systems, Knowledge Graphs (KGs) offer a structured, semantically rich framework that has a long-standing history of enhancing fact-checking efforts. KGs play a crucial role in encapsulating and organizing complex information through their inherent structure which is comprised of triples. Each triple, consisting of a subject, predicate, and object — al-

ternatively framed as a head entity, a relation, and a tail entity *i.e.*, (subject entity, relationship, object entity) — constitutes the core component of a KG, enabling it to represent structural facts and support symbolic reasoning effectively.

KGs represent data in a way that captures information about not just the entities but also the complex relationships between them. This semantic web of information allows for a deeper understanding of context, which is essential for verifying facts. Furthermore, KGs facilitate the exploration of multi-hop information pathways, allowing for the elucidation of intricate and indirect relationships critical for comprehensive fact verification. Prior work has shown promising results utilizing KGs (Hu et al., 2023; Liu et al., 2020b; Ma et al., 2023). However, concurrently integrating both the structured knowledge graphs with unstructured text as inputs to LLMs is not a trivial enterprise. Prior work has tried directly including triples as input to LLMs (Baek et al., 2023; Sequeda et al., 2023). Yet LLMs are not trained for leveraging triples, and this approach does not leverage the community and entity relationship. Other approaches (Sun et al., 2021; Liu et al., 2020a; Yasunaga et al., 2022; Sun et al., 2020; Zhang et al., 2022; Kang et al., 2023) require training customized models or joint embeddings that are computationally expensive.

In light of the distinct advantages of KGs and the capabilities of RAG systems and LLMs, the absence of research on their combined application for fact-checking is notable. Although such integration —— melding KGs' structured, semantic insights with RAG's dynamic retrieval and LLMs' language comprehension —— holds significant promise for advancing fact-checking technologies, the specific impact of this synergistic approach remains largely unexplored.

To bridge the existing research gap, we introduce a pioneering framework: **CommunityKG-RAG (Community Knowledge Graph-Retrieval Augmented Generation)**. This innovative approach synergizes Knowledge Graphs with Retrieval-Augmented Generation and Large Language Models to enhance fact-checking capabilities. By leveraging and preserving the intricate entity relationships and community structures within KGs, our framework provides a contextually enriched and semantically aware retrieval mechanism that significantly improves the accuracy and relevance of generated responses. Specifically, we construct a comprehensive KG from fact-checking articles, employ the Louvain algorithm for community detection, and assign embeddings derived from word embeddings to each node. This approach ensures that the identified communities are both structurally coherent within the KG and highly pertinent to the fact-checking task. By harnessing this integrated framework, we offer a robust, scalable, and efficient solution to contemporary fact-checking challenges. An example of this integration and its impact on retrieval accuracy is illustrated in Figure 1.

Our contributions are threefold:

1. **Utilization of Both Structured and Unstructured Data with Superior Knowledge Graph Integration:** By combining the structured data of Knowledge Graphs with the unstructured data handled by LLMs, we achieve a more comprehensive and context-aware fact-checking system. We demonstrate that converting knowledge graphs back to sentences within our framework is superior to methods that use triples as context. This approach enhances the comprehensibility and relevance of the retrieved information, as demonstrated by the significant increase in accuracy.

2. **Context-Aware Retrieval and Multi-hop Utilization:** By leveraging community structures and multi-hop paths within KGs, the framework delivers more precise and relevant information retrieval, enhancing the overall effectiveness of the fact-checking process. We are the first work to propose utilizing and combining multi-hop in KGs with RAG systems.

3. **Scalability and Efficiency:** The framework operates in a zero-shot manner, requiring no additional training or fine-tuning, which ensures high scalability and adaptability to various LLMs. Additionally, the knowledge graph and community detection processes only need to be performed once, allowing for repeated reuse or rapid updates.

2 Related Work

KGs in LLM inputs

Recent research has explored the integration of KGs with LLMs, where triples are directly fed into LLMs as input (Baek et al., 2023; Sequeda et al., 2023). However, this approach has its limitations, particularly in its assumption that LLMs

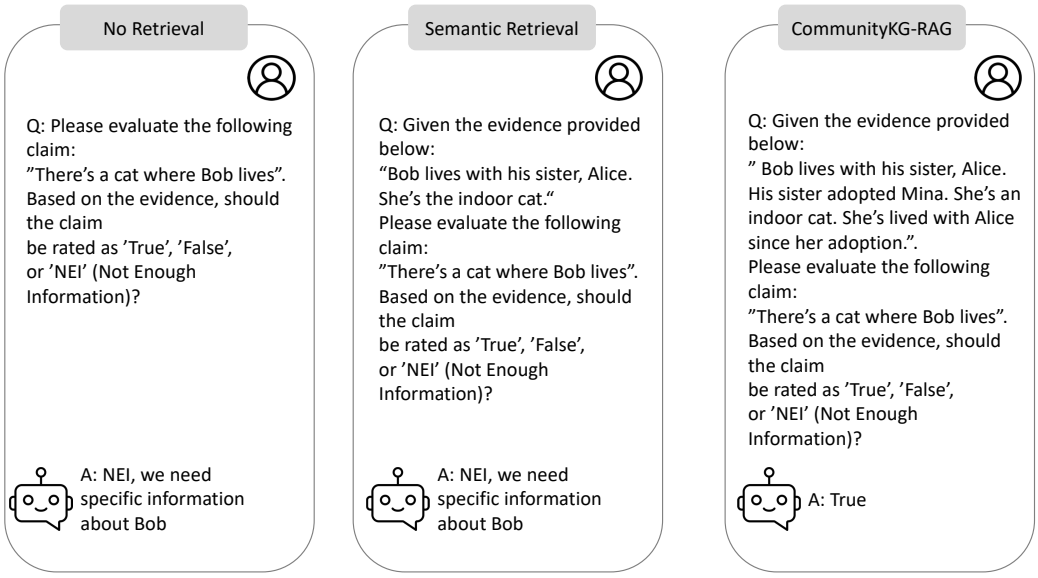

Figure 1: Comparison between no retrieval, semantic retrieval, and CommunityKG-RAG. The no retrieval and semantic retrieval fail to provide sufficient context, while our proposed method, CommunityKG-RAG, is able to by leveraging multi-hop knowledge graph information in the retrieval process enhancing accuracy and relevance.

can effectively process and utilize triples despite their primary training focus on sequential data processing. This could result in an underutilization of KG's structural information, such as subgraph structure, community structure, and relationship patterns across entities and relations of Knowledge Graphs. Addressing this, our proposed method leverages community detection results as indices for text retrieval, thus harnessing the subgraph and entity relationship structures inherent in KGs more effectively than in previous work.

Other approaches to integrating knowledge graphs with language models include joint embedding training or the customization of model architectures. This can be done by representing triplets as a sequence of tokens and concatenating them with text embedding in the pre-training stage (Sun et al., 2021; Liu et al., 2020a). For instance, Yasunaga et al. (2022) propose a cross-modal model to fuse text and KG to jointly pre-train the model. Sun et al. (2020) present a word-knowledge graph that unifies words and knowledge. Zhang et al. (2022) fuses representations from pre-trained language models and graph neural networks over multiple layers. Models that require additional training are computationally expensive and cumbersome. Kang et al. (2023) retrieves a relevant subgraph composed of triples by utilizing GNN for triple embedding. In contrast, our method does not necessitate additional training, offering a more efficient and adaptable solution for integrating KGs with LLMs.

3 Problem Statement

The goal of fact-checking task formulation is to locate the top n most relevant sentence, in order to classify a given claim as either *refuted*, *supported*, or *not enough information* as the labels by a large language model. Let P represent a corpus of fact-checking articles and C a set of claims. Each claim $c \in C$ is associated with a ground-truth label y. There exists a set of top k most relevant sentences $P_c = p_i^k$ from the fact-checking articles P for each claim c. The task is formulated as optimizing the prediction $\hat{y} = f(C, P_c)$, where f is a large language model to evaluate the truthfulness of claims based on the evidence provided.

4 CommunityKG-RAG

In this section, we detail our novel framework CommunityKG-RAG for integrating KGs with RAG systems and LLMs to enhance fact-checking capabilities. We show an overview in Figure 2. Our approach leverages the structural advantages of KGs to provide a contextually enriched, semantically aware information retrieval mechanism, which is subsequently used to inform the generation process of LLMs.

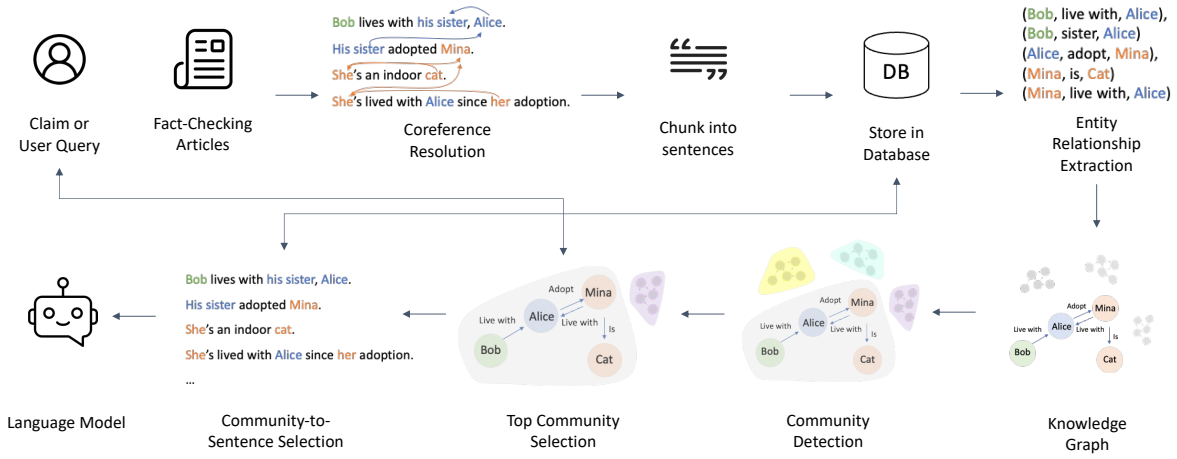

Figure 2: Workflow of CommunityKG-RAG

4.1 Knowledge Graph Construction

We begin by constructing a KG from a corpus of fact-checking articles. The construction process involves the following three steps:

4.1.1 Coreference Resolution

Coreference resolution is a preprocessing step to enhance the semantic coherence of the input data prior to knowledge graph construction. This process aims to identify and cluster mentions of entities and pronouns that refer to the same real-world entities across the corpus, thereby resolving ambiguities in entity references.

We employ a state-of-the-art coreference resolution model by Lee et al. (2018), leveraging a deep learning approach based on SpanBERT (Joshi et al., 2020), which has been pre-trained on a large corpus to capture a wide range of syntactic and semantic information.

4.1.2 Graph Construction

CommunityKG-RAG leverages the relationship extraction model, REBEL, proposed by Cabot and Navigli (2021) to discern entity relationships within the corpus. This process is formalized as follows:

Given the corpus P, we extract a set of entities, denoted as $E = \{e_1, e_2, ..., e_n\}$. We construct the entity graph $G = (E, R)$, where R comprises the set of relationships between entities. In this graph, entities (E) are represented as nodes, and relationships (R) are depicted as edges that link these nodes. This graph represents the intricate network of connections among entities derived from the corpus, forming the foundation of the KG.

This structured approach facilitates a comprehensive representation of the factual relationships within articles, thereby enabling advanced analysis and application in fact-checking and misinformation identification tasks.

4.1.3 Node Feature Embedding

For each node in the KG, we assign it with word embeddings derived from a pre-trained BERT model (Devlin et al., 2018a). This embedding serves as the node feature vector, encapsulating the semantic information of the entity.

4.2 Community Detection

To leverage the community structures inherent within the Knowledge Graph (KG) for enhanced fact retrieval, we employ the Louvain algorithm (Blondel et al., 2008) as a foundational tool. This algorithm is instrumental in detecting and delineating communities within the graph G, by focusing on the optimization of modularity. Modularity is a scalar value between -1 and 1 that measures the density of links inside communities compared to links between communities. The algorithm initially treats each node as its own community and itera-

tively merges communities to maximize the gain in modularity. This optimization continues until no further improvement in modularity is possible, resulting in a partition of the graph into distinct communities.

From graph G, we extract a set of communities denoted by M, where each community $m \in M$ represents a cluster of nodes more interconnected among themselves than with the rest of the graph. This structured approach allows us to focus our retrieval efforts on specific segments of the KG that are more likely to contain relevant and contextually rich information for fact-checking tasks.

4.3 Community Retrieval

Each community m is considered as a subgraph $G_m = (E_m, R_m)$ comprising a subset of entity nodes E_m and their relationships R_m. The embedding representation of each community denoted as $\varphi(m)$ is derived by averaging the BERT embeddings of the nodes within E_m:

$$\varphi(m) = \frac{1}{|E_m|} \sum_{i \in E_m} \text{BERT}(e_i)$$

where $|E_m|$ is the number of nodes in a community m and e_i represents the word embedding of node i derived from BERT model (Li et al., 2023a) as described in the section 4.1.3. This approach aggregates the collective semantic attributes of the community, encapsulating a comprehensive semantic representation.

To convert claims into embeddings for similarity comparisons, we utilize the BERT-base Sentence Transformer model, Sentence-BERT (Reimers and Gurevych, 2019). Sentence-BERT is specifically optimized for generating high-quality sentence embeddings, making it ideally suited for comparing the semantic similarities between claims and community descriptions.

The relevance score $r(c, m)$ between claim c and community m is calculated as the dot product between their embeddings:

$$r(c, m) = \varphi(c)^T \varphi(m)$$

4.4 Top Community Selection

To efficiently prioritize communities for deeper analysis, the top δ percent of communities, ranked by their relevance scores $r(c, m)$, are selected. The selection threshold N is determined as follows: $N = \lceil \frac{\delta}{100} \times |M| \rceil$, where $|M|$ represents the total number of communities. Consequently, the subset of most relevant communities M_c^* to claim c is defined as:

$$M_c^* = \{m \in M : \text{rank}(r(c, m)) \leq N\}$$

This selection criterion ensures that our analysis is concentrated on the communities most likely to contain relevant and substantive information pertinent to the verification of a claim c, thus facilitating efficient and focused fact-checking.

4.5 Top Community-to-Sentence Selection

To identify the most pertinent sentences, a relevance score $r(M_c^*, p)$ is computed for each sentence p within the top communities M_c^*. Sentences are then ranked by relevance, and the top λ percent are selected, resulting in a subset P_c^* of the most relevant sentences.

This structured approach allows for systematic filtering and selection of significant information, a process which is crucial for robust and focused fact-checking. We use CommunityKG-RAG$_\lambda^\delta$ to represent the synergistic application of two distinct filters: the top δ percent for community relevance and the top λ percent for sentence significance within the context of validating community-to-sentence relevance. This refined designation underscores a strategic methodological synthesis aimed at optimizing the fact-checking process by methodically concentrating on the most pivotal communities and their essential corresponding sentences.

5 Experimental Details

5.1 Datasets

MOCHEG This multimodal fack-checking dataset (Menglong Yao et al., 2022) consists of 15,601 claims annotated with a truthfulness label collected from PolitiFact and Snopes, two popular websites for fact-checking articles. The articles and results of claim verification were produced by journalists manually. The truthfulness is labeled into three categories: supported, refuted, and NEI (not enough information). More details are included in the Appendix A.

5.2 Baselines

No Retrieval This is a naive baseline where answers are generated from the language model through prompts without context or retrieval.

Semantic Retrieval Following Nie et al. (2019), we extract context based on semantic similarity. Specifically, we use cosine similarity in embeddings between the prompt and the context. BERT (Devlin et al., 2018b) is used to produce the embedding.

Knowledge-Augmented language model PromptING (KAPING) We implement KAPING proposed by Baek et al. (2023). The KAPING is a zero-shot RAG framework that proposes basing retrieval on sentence similarity between the input text and triples. The output prompt of the KAPING framework includes the original text prompt with triples as the context. Specifically, the triples are in the format of $(subject entity, relationship, object entity)$. We equip KAPING with the same set of articles for retrieval.

5.3 Implementation Details

We conducted our experiments using the LLaMa2 7 billion model as our primary Large Language Model (Touvron et al., 2023). The LLaMa2 models are open-source and widely accessible. We chose these models because they were trained on trillions of tokens, including publicly available datasets like Wikipedia, and demonstrated state-of-the-art results at the time when the texts were published. This capability enabled a thorough evaluation of our method's zero-shot performance when applied to previously unseen corpora.

The availability of these models in multiple sizes enabled a comparative analysis of our proposed framework, assessing how model scale impacts performance. Furthermore, since Wikipedia was integral to their training datasets, we were able to explore the efficacy of our approach on corpora familiar to the models. The utility of this retrieval approach has been substantiated in prior research (Khandelwal et al., 2020).

To quantitatively assess the LLMs, we measured their performance in verifying claims using accuracy as our metric. More details of the LLMs and the corresponding prompt are included in Appendices B and C.

We use CommunityKG-RAG$^{25}_{100}$ as the baseline. In other words, we use the top $\delta = 25$ percent of the most relevant communities and $\lambda = 100$ percent of the sentences that the community maps to as the context.

Model	LLaMa2 7B
No Retrieval	39.79%
Semantic Retrieval	43.84 %
KAPING	39.41 %
CommunityKG-RAG$^{25}_{100}$	**56.24%**

Table 1: Comparison of claim verification accuracy for various retrieval methods: No Retrieval, Semantic Retrieval, KAPING, and our approach, CommunityKG-RAG$^{25}_{100}$, which selects the top 25 percent of relevant communities and uses 100 percent of their mapped sentences as context.

6 Results

6.1 Main Results

Overall, our proposed method, CommunityKG-RAG$^{25}_{100}$, not only achieves the best results but also surpasses all baselines, as detailed in Table 1. The No Retrieval baseline recorded an accuracy of 39.79 percent. Employing the Semantic Retrieval strategy yielded an improvement, elevating accuracy to 43.84 percent. This increase underscores the advantages of integrating semantic context, thereby enhancing the proficiency of the language model in claim verification.

Conversely, the KAPING method did not enhance performance, registering a slight decline in accuracy to 39.41 percent. This outcome indicates that a language model such as LLaMa2 may struggle with retrieval contexts formatted as triples (*i.e.*, (subject entity, relationship, object entity)). Such structuring appears to impede the model's capacity to effectively utilize information. This is likely due to its foundational training on sequential word prediction rather than on processing structured data like triples.

However, the performance of our approach, CommunityKG-RAG$^{25}_{100}$, was markedly superior, achieving an accuracy of 56.24 percent. This significant increase not only confirms the effectiveness of integrating community-derived knowledge into the retrieval process but also demonstrates substantial gains over conventional retrieval methods. These results validate the substantial impact that tailored, community-focused retrieval mechanisms can have on the operational effectiveness of language models in complex verification scenarios. This marked improvement reiterates the critical role of precise, context-aware retrieval strategies in augmenting the functional capabilities of language models.

6.2 Ablation

We conducted a series of ablation studies to understand the significance of various factors within the CommunityKG-RAG framework. Specifically, these ablation studies are designed to evaluate the impact of different backbone language models, the selection of top communities, and the extent of community-to-sentence selection.

6.2.1 Performance With Different Backbone Models

To demonstrate the robustness and adaptability of the proposed CommunityKG-RAG framework, we conducted an ablation study to assess how different backbone language models affect the performance on the MOCHEG fact-checking dataset. Considering the computational costs, which increase with the number of communities and community-to-sentences selection using the community (Appendix E), we conduct this ablation with CommunityKG-RAG$^{25}_{25}$. We selected the top $\delta = 25$ percent of the most relevant communities and the top $\lambda = 25$ percent of the sentences mapped to these communities to serve as the contextual input.

In this analysis, we compared the performance of two different backbone models: LLaMa2 7B and LLaMa3 8B. Table 2 illustrates the outcomes, showing that CommunityKG-RAG significantly enhances performance across both models. Specifically, when employing the CommunityKG-RAG framework, there is a notable improvement of 6.18 percentage points with LLaMa2 7B and an increase of 3.21 percentage points with LLaMa3 8B compared to the no retrieval baseline. However, we observed that the LLaMa3 8B showed a lesser improvement and accuracy over the no retrieval baseline than the 7B model despite its larger size. This may be attributed to the 8B model's capability to explore various facets of a given issue more comprehensively, which, while generally beneficial, might lead to a less precise matching in scenarios demanding exact binary evaluations, such as our fact-checking tasks. This characteristic could also contribute to the slightly lower improvement observed with the 8B model.

These results underscore the effectiveness of our framework in leveraging structured community knowledge, thereby improving the accuracy of fact-checking across diverse language model architectures.

Model	**LLaMa2 7B**	**LLaMa3 8B**
No Retrieval	39.79%	26.03%
CommunityKG-RAG$^{25}_{25}$	45.97%	29.24%

Table 2: Performance comparison of no retrieval and CommunityKG-RAG with $\delta = 25$ and $\lambda = 25$ settings across different backbone models, LLaMa2 7B and LLaMa3 8B.

6.2.2 Influence of Community-to-Sentence Selection

This section examines the influence of varying community-to-sentence selection thresholds within a consistently held community threshold of 25 percent on the performance of the CommunityKG-RAG framework using the LLaMa2 7B model. Community-to-sentence selection thresholds were adjusted to 25 percent, 50 percent, 75 percent, and 100 percent to identify the optimal level for enhancing fact-checking performance.

Model	LLaMa2 7B
CommunityKG-RAG$^{25}_{25}$	45.97%
CommunityKG-RAG$^{25}_{50}$	27.83%
CommunityKG-RAG$^{25}_{75}$	41.93%
CommunityKG-RAG$^{25}_{100}$	56.24%

Table 3: Performance variations of the LLaMa2 7B model under the CommunityKG-RAG framework with consistent community threshold (top 25 percent) and variable community-to-sentence selection.

The results presented in Table 3 demonstrate variable model performance as community-to-sentence selection thresholds change. Initially, the performance slightly decreases to 27.83 percent when the inclusion rate of sentences is increased from 25 percent to 50 percent. This might indicate that the top 25 percent of sentences contain the most crucial information for verifying the claim, and including additional sentences up to 50 percent introduces noise or less relevant data that temporarily hinder the model's accuracy. However, as the inclusion rate continues to increase to 75 percent and then to 100 percent, the performance improves, ultimately achieving the highest accuracy at a full 100 percent inclusion rate. This suggests that beyond the 50 percent threshold, the additional sentences contribute positively, possibly by providing necessary context that supports more accurate fact-checking.

This pattern highlights the critical role of exten-

sive contextual engagement in the CommunityKG-RAG framework, demonstrating that access to a wider array of sentences associated with a carefully selected group of communities markedly improves the model's effectiveness in accurately identifying truth and falsehood. These results underscore the nuanced balance needed in selection strategies to provide adequate context for accurate analysis without inundating the model with extraneous data.

6.2.3 Combined Effects of Top Community and Community-to-Sentence Selection

To further explore the efficacy of the CommunityKG-RAG framework, we conducted an analysis to understand the impact of varying the top community and community-to-sentence selection thresholds on the performance of the model. We adjusted the thresholds of both δ and λ to 25 percent, 50 percent, 75 percent, and 100 percent to examine how the extent of considered context in both community and community-to-sentence selection affect the fact-checking capabilities of the CommunityKG-RAG framework. We show the knowledge graph community statistics in Appendix E.

The results, as shown in Table 4, reveal interesting trends. Initially, the increase of thresholds from 25 percent to 75 percent led to a slight decrease in performance, suggesting that adding more communities and sentences might introduce noise or less relevant information, thus compromising the model's effectiveness. However, a significant improvement is observed when the thresholds are expanded to 100 percent. This enhancement at the highest threshold suggests that the model benefits from a more comprehensive view of the available data, possibly capturing essential contextual nuances that are otherwise missed at lower thresholds. This pattern aligns with observations from previous ablation studies concerning community-to-sentence selection.

Interestingly, when comparing the effects of top community selection, an increase in the number of top communities results in improved accuracy while holding community-to-sentence selection constant. This observation emerges from comparing CommunityKG-RAG$^{25}_{50}$ versus CommunityKG-RAG$^{50}_{50}$, and CommunityKG-RAG$^{25}_{75}$ to CommunityKG-RAG$^{75}_{75}$.

However, increasing both the community selection and community-to-sentence selection to 100 percent does not yield further improvements. As illustrated by the comparison between CommunityKG-RAG$^{25}_{100}$ and CommunityKG-RAG$^{100}_{100}$, this finding implies that a targeted selection of highly relevant communities, along with a comprehensive examination of their associated sentences, strikes an ideal balance. It enables the model to access detailed and pertinent information effectively without being overwhelmed by extraneous data. This method provides a nuanced approach to information retrieval that maximizes accuracy while avoiding information overload.

Model	LLaMa2 7B
CommunityKG-RAG$^{25}_{25}$	45.97%
CommunityKG-RAG$^{50}_{50}$	43.64%
CommunityKG-RAG$^{75}_{75}$	43.60%
CommunityKG-RAG$^{100}_{100}$	54.62%

Table 4: Performance metrics of the LLaMa2 7B model within the CommunityKG-RAG framework across varied thresholds of top community and community-to-sentence selection. The table details the model's accuracy percentages at incremental selection thresholds of 25, 50, 75, and 100 percent for both community and community-to-sentence selection, illustrating how varying levels of context inclusion impact the model's performance.

7 Conclusion

We have introduced CommunityKG-RAG, a novel framework that integrates Knowledge Graphs with Retrieval-Augmented Generation and Large Language Models to enhance fact-checking. This approach leverages the structured data of KGs and the generative capabilities of LLMs, significantly improving the accuracy and relevance of responses.

CommunityKG-RAG effectively addresses key challenges such as outdated information and hallucinations by utilizing multi-hop community structures for refined and accurate retrieval within KGs. This integration enables more precise and contextually rich information retrieval, crucial for effective fact-checking. Our framework achieves superior performance without requiring any fine-tuning or additional training, demonstrating its robustness and efficiency. As the first framework to combine multi-hop community information in KGs with RAG systems, CommunityKG-RAG represents a significant advancement and promising direction for future work.

8 Limitations

Despite the notable success of the CommunityKG-RAG framework in enhancing claim verification accuracy, several limitations highlight areas for future research and improvement:

8.1 Computational Demands

The CommunityKG-RAG framework places substantial demands on computational resources compared to no retrieval or semantic retrieval. However, communities can be pre-computed and reused, making the operational phase more lightweight and dynamic. This capability enhances the model's responsiveness to new data and trends. Further, our method has demonstrated significant accuracy improvements despite the computational demands, and, besides, our proposed method is more lightweight than methods that require training or fine-tuning a language model.

8.2 Dependency on Entity Recognition Quality

Our proposed method's effectiveness heavily relies on the quality of entity recognition. There are prior works (Edge et al., 2024) that rely on utilizing language models to conduct entity recognition. This could potentially introduce hallucinations. To avoid such risk, we use REBEL, a seq2seq model based on Wikipedia data. If the framework is applied to text that is significantly different from Wikipedia text, it might hinder performance. In such cases, utilizing an entity recognition method tailored to the specific domain could be beneficial. However, as shown in the Appendix E, our approach incorporates a comprehensive dataset with up to 48,630 nodes and 202,455 edges, which ensures a robust and extensive knowledge base. This comprehensive coverage helps mitigate potential quality issues, enhancing the reliability of the entity recognition process.

These limitations, alongside the outlined implementation advantages, underscore the need for ongoing refinement and testing of the CommunityKG-RAG framework to optimize its practicality and effectiveness in real-world scenarios. The ability to pre-compute communities ensures that the method remains operationally lightweight and scalable, an essential factor for broad application. Additionally, future work can consider extending this method framework into multimodality, integrating multimodal graphs or tabular data. Such extensions could further enhance the model's capabilities and applicability in more complex and varied data environments, opening new avenues for research and practical implementation.

References

Jinheon Baek, Alham Fikri Aji, and Amir Saffari. 2023. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. *arXiv preprint arXiv:2306.04136*.

Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, and Mohamed Abdelrazek. 2024. Seven failure points when engineering a retrieval augmented generation system. *arXiv preprint arXiv:2401.05856*.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Mars Gokturk Buchholz. 2023. Assessing the effectiveness of gpt-3 in detecting false political statements: A case study on the liar dataset. *arXiv preprint arXiv:2306.08190*.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. Rebel: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381.

Kevin Matthe Caramancion. 2023. Harnessing the power of chatgpt to decimate mis/disinformation: Using chatgpt for fake news detection. In *2023 IEEE World AI IoT Congress (AIIoT)*, pages 0042–0046. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018a. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018b. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization.

Emma Hoes, Sacha Altay, and Juan Bermeo. 2023. Leveraging chatgpt for efficient fact-checking. *PsyArXiv. April*, 3.

Xuming Hu, Junzhe Chen, Zhijiang Guo, and Philip S. Yu. 2023. Give me more details: Improving fact-checking with latent retrieval. *arXiv preprint arXiv:2305.16128*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.

Yue Huang and Lichao Sun. 2023. Harnessing the power of chatgpt in fake news: An in-depth exploration in generation, detection and explanation. *arXiv preprint arXiv:2310.05046*.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Minki Kang, Jin Myung Kwak, Jinheon Baek, and Sung Ju Hwang. 2023. Knowledge graph-augmented language models for knowledge-grounded dialogue generation.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Miaoran Li, Baolin Peng, and Zhu Zhang. 2023b. Self-checker: Plug-and-play modules for fact-checking with large language models. *arXiv preprint arXiv:2305.14623*.

Hao Liao, Jiahao Peng, Zhanyi Huang, Wei Zhang, Guanghua Li, Kai Shu, and Xing Xie. 2023. Muser: A multi-step evidence retrieval enhancement framework for fake news detection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020a. K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2901–2908.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020b. Fine-grained fact verification with kernel graph attention network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7342–7351, Online. Association for Computational Linguistics.

Jing Ma, Chen Chen, Chunyan Hou, and Xiaojie Yuan. 2023. KAPALM: Knowledge grAPh enhAnced language models for fake news detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3999–4009, Singapore. Association for Computational Linguistics.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2022. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. *arXiv e-prints*, pages arXiv–2205.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6859–6866.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Juan Sequeda, Dean Allemang, and Bryon Jacob. 2023. A benchmark to understand the role of knowledge graphs on large language model's accuracy for question answering on enterprise sql databases. *arXiv preprint arXiv:2311.07509*.

Amir Soleimani, Christof Monz, and Marcel Worring. 2019. Bert for evidence retrieval and claim verification.

Tianxiang Sun, Yunfan Shao, Xipeng Qiu, Qipeng Guo, Yaru Hu, Xuanjing Huang, and Zheng Zhang. 2020. Colake: Contextualized language and knowledge embedding. *arXiv preprint arXiv:2010.00309*.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, Weixin Liu, Zhihua Wu, Weibao Gong, Jianzhong Liang, Zhizhou Shang, Peng Sun, Wei Liu, Xuan Ouyang, Dianhai Yu, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Yile Wang, Peng Li, Maosong Sun, and Yang Liu. 2023. Self-knowledge guided retrieval augmentation for large language models. *arXiv preprint arXiv:2310.05002*.

Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D Manning, Percy S Liang, and Jure Leskovec. 2022. Deep bidirectional language-knowledge graph pretraining. *Advances in Neural Information Processing Systems*, 35:37309–37323.

Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D Manning, and Jure Leskovec. 2022. Greaselm: Graph reasoning enhanced language models for question answering. *arXiv preprint arXiv:2201.08860*.

A Details of Datasets

The dataset was partitioned into training and testing subsets, with the training set employed for constructing the knowledge graph and verifying claim accuracy. Comprising 18,553 unique claims, each is linked to a corresponding fact-checking article and label.

The target variable, "truthfulness," is classified into three categories: "Supported," "Refuted," and "Not Enough Information" (NEI). The label distribution includes 7,137 "Refuted," 6,928 "Supported," and 4,488 "NEI."

Label assignment for "Supported," "Refuted," and "NEI" was performed following a meticulous cleaning process carried out by the authors of Menglong Yao et al. (2022). This process was conducted as the original labels derived from the fact-checking articles were marred by noise and inconsistency. Initially, the labels encompassed a broad spectrum of classifications, including "False," "Mostly False," and "Half True," totaling up to 75 different labels. This refinement was crucial as the original articles did not explicitly categorize claims into "Supported," "Refuted," or "NEI." This ambiguity could potentially impair the retrieval capabilities of large language models (LLMs). To mitigate this, we simplified the labels by mapping "Supported" to "True" and "Refuted" to "False" during the prompting and preprocessing phases.

B Prompt

The prompt used for all RAG systems is the following:

```
"Given the evidence provided below:
{formatted_evidence}.
Please evaluate the following claim:
{claim}.
Based on the evidence, should the claim
be rated as 'True', 'False',
or 'NEI' (Not Enough Information)?"
```

The prompt used for all baseline zero shot setups is the following:

```
"Please evaluate the following claim:
{claim}.
Based on the evidence, should the claim
be rated as 'True', 'False',
or 'NEI' (Not Enough Information)?"
```

C Language Model Parameters

In our experiments, we utilized the meta-llama/Llama-2-7b-chat-hf model from Hugging Face's model hub. Our generation pipeline was configured to produce coherent and non-repetitive text. Key settings included a temperature of 0.3 to encourage predictability, a repetition penalty of 1.1 to avoid redundant content, and a limit of 200 new tokens per output to maintain focus. Custom stopping criteria were implemented to end text generation at specific tokens, ensuring outputs remained within the scope of our conversational framework.

D Computing Infrastructure

All computational experiments were conducted on a server configured with two NVIDIA RTX A6000 GPUs, each with 48 GB of GDDR6 memory, and two AMD EPYC 7513 32-core processors. The system also included 512 GB of DDR4 ECC RAM and a 960 GB Samsung PM983 NVMe SSD for storage.

E Community Statistics

We provide the knowledge graph community statistics with various top δ percent communities in Table 5. These statistics demonstrate the multi-hop nature of our knowledge graphs through the metrics of average shortest path length and diameter. The average shortest path length, ranging from 4.03 to 4.28 across different community percentages, indicates that on average, multiple hops are required to traverse between nodes. The diameter values, ranging from 13 to 17, suggest the presence of long paths within the graphs, further supporting the existence of multi-hop pathways. These metrics confirm that our CommunityKG-RAG framework effectively leverages multi-hop connections, crucial for retrieving contextually rich and relevant information in fact-checking tasks.

Metric	Value
	Top 25 Percent
Number of Nodes	20,092
Number of Edges	60,770
Avg. Degree	6.05
Avg. Communities per Claim	2.05
Avg. Nodes per Claim	5.62
Avg. Shortest Path Length	4.28
Diameter	17
	Top 50 Percent
Number of Nodes	32,428
Number of Edges	117,677
Avg. Degree	7.26
Avg. Communities per Claim	4.57
Avg. Nodes per Claim	11.63
Avg. Shortest Path Length	4.13
Diameter	13
	Top 75 Percent
Number of Nodes	40,669
Number of Edges	159,703
Avg. Degree	7.85
Avg. Communities per Claim	6.85
Avg. Nodes per Claim	16.60
Avg. Shortest Path Length	4.07
Diameter	14
	Top 100 Percent
Number of Nodes	48,630
Number of Edges	202,455
Avg. Degree	8.33
Avg. Communities per Claim	9.64
Avg. Nodes per Claim	22.25
Avg. Shortest Path Length	4.03
Diameter	13

Table 5: Community Statistics